Evaluating data-driven style transformation for gesturing embodied agents

Alexis Héloir, Michael Kipp, Sylvie Gibet, Nicolas Courty

Evaluating data-driven style transformation for gesturing embodied agents

Alexis Heloir[*], Michael Kipp[*], Sylvie Gibet[**] and Nicolas Courty[**]

[*]DFKI - German Research Center for Artificial Intelligence, Campus D3.2, 66123 Saarbrücken, Germany
[**] Laboratoire Valoria - Université de Bretagne Sud Campus de Tohannic, 56000 Vannes,
[*]firstname.surname@dfki.de, [**]firstname.surname@univ-ubs.fr

Abstract. This paper presents an empirical evaluation of a method called "Style transformation" which consists of modifying an existing gesture sequence in order to obtain a new style where the transformation parameters have been extracted from an existing captured sequence. This data-driven method can be used either to enhance key-framed gesture animations or to taint captured motion sequences according to a desired style.

1 Introduction

Endowing a virtual humanoid with expressive gestures requires one to take into account the properties that influence the perception of convincing movements. One of these properties can be called style. Style gathers all subtle charateristics occuring both over spatial and temporal aspects of motion. Style also gives information about the speaker's age, gender, cultural background, and emotional state. As a consequence, style contributes to making a virtual humanoid more convincing which makes it more acceptable to human useres. This paper presents an empirical evaluation of a method called "Style Transformation" which consists of applying an automatic style on a neutral input motion to generate an appropriate style variant.

Style transformation basically consists of modifying an existing gesture sequence in order to obtain a new style whose transformation parameters have been extracted from an existing captured sequence. This data-driven method can be used either to enhance artist-generated gesture animations or to taint captured motion sequences according to a desired style.

2 Related work

Work dedicated to the specification and generation of expressive gestures can be separated into two categories: gesture selection (which gesture is most suitable to be displayed) and motion quality (how should the gesture be displayed). In this section, we focus on the motion quality. Again, motion-quality dedicated work

may be separated into two parts: theory driven approaches and data driven approaches.

Theory driven apporaches are driven by expert knowledge gained from empirical studies and extensive observation of human motion. the underlying history derived from foundation work [1, 2] which served as a base for procedural motion synthesis systems [3, 4]. Procedural systems have been proved to be capable of providing understandable expressive gestures with a high level of control [5, 6]. However, such generative models, by relying on kinematics and/or physical models, have failed to produce natural motions. Such a lack motivates our investigation towards data driven approaches.

Recently, expressivity dedicated studies presented empirical implementations of expressivity models [7, 8, 5]. While Buisine et al. [8] addressed the perception of expressive features and their influence over generated agent's perceived behaviors (briskness, wearyness, tonicity), Schröder [7] studied the relationship between the 3 dimensions of emotion (valence, arousal, dominance) and the perceived emotion for speech synthesis. Kipp et al. [5] performed an empirical study dedicated to the influence of gesture-unit length over "believability" and "sympathy" of the produced gestures. Our experimental study is inspired by this prior work, as we present an empirical evaluation of generated motion and establish comparisons on the relation between perceived gesturing style and the valence, arousal, dominance (VAD) model of emotions.

3 Style transformation overview

This section presents an overview of the style transformation pipeline. The style transformation can be decomposed into two stages. A modeling stage where relevant parameters of the transformation model are inferred from existing data and a transformation stage where an arbirary input motion is transformed to convey the style which has been inferred during the learning stage. Figure 3 depicts the different stages involved in the style transformation pipeline. In the following paragraphs is a presentation of the gesture material which has been used as an input to our method followed by a short overview of the style tranformation method.

The experiments presented in this paper rely on four motion captured French Sign Language (FSL) gesture sequences performed by a deaf professional instructor. Three of these gesture sequences depict a weather forecast presentation performed according to different styles the signer had been asked to mimic: neutral (*-n*), angry (*-a*) and weary (*-w*). The fourth sequence depicts information usually displayed in railway stations, this sequence was performed according to a neutral style. In the following, the three styled gesture sequences (neutral, angry and weary) from the weather forecast material will be refered to as (*M-W-n*, *M-W-w*, *M-W-a*) while the sequence depicting train information will be refered to as (*M-T-n*). Motion data has been acquired using a method described in [9].

One way of conveying subtle aspects of gesture style for articulated figures is to take into account the dependencies between joint motions. The spatial trans-

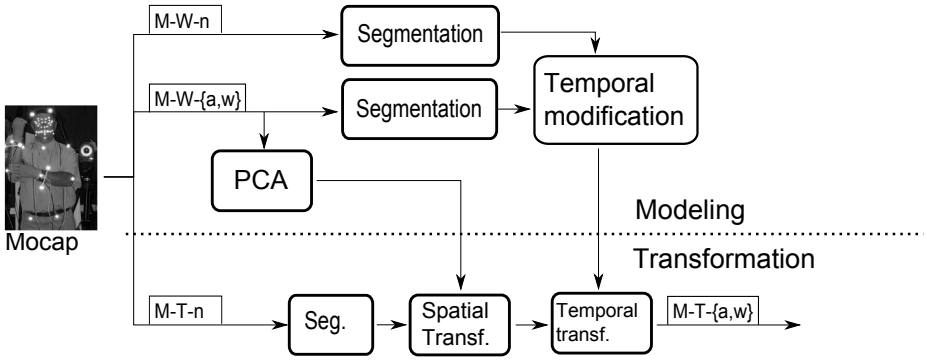

Fig. 1. Overview of the style transformation pipeline

formation introduced in this paper relies on the principal component analysis (PCA) and is comparable to the Egges et al. proposal [10]. However, this method is applied to the whole body structure and spatial corrections are then applied using inverse kinematics afterwards to prevent undesired foot skating effect.

Style transformation can be achieved on the temporal dimension using motion retiming. Although non-linear playback-speed modification is straightforward, determination of a relevant time-deformation profile may be tedious. A more in-depth explanation of the method is given in [11, 12].

4 Perceptive evaluation of style transformation

This section presents the empirical evaluation which has been conducted in order to test the accuracy of our style transformation method and check how far users are able to discriminate between the three styles using questions along three dimensions.

4.1 Experimental setup

Participants The experiment took the form of a questionnaire. Although the questionnaire was answered using an on-line web interface, the study was restricted to 19 subjects so that it could be better controlled. Subjects were German-speaking students from Saarbrucken University, Germany. Participants were aged between 20 and 25. Most of them (17 out of 19) were psychology or CS. Students' participation in such an experiment is required in the context of their curriculum. Experiment data has been gathered anonymously.

Method Each evaluation was performed in our lab and was supervised by a member of our team. Every subject passed the evaluation individually. The test consisted of visualizing 9 video-enabled questionnaires and each time comparing

3 styled realizations of a gesture clip. The video embedded in the questionnaire depicted short clips (mostly phrase segments) of motion captured weather forecast sequences (5 clips × 3 styles: *M-W-n*, *M-W-a*, *M-W-w*) and train information sequences (4 clips × 3 styles, 1 captured style: *M-T-n* and two generated styles: *M-T-a*, *M-T-w*) as depicted in section 3. 27 clips have been generated in total. All clips have been rendered in the same manner: agent appearance, lighting conditions, camera position and orientation were the same for each gesture clip. Rendering of our generated motion files was perfomed in Blender's [1] internal rendering engine. A questionnaire is depicted in Figure 4.1.

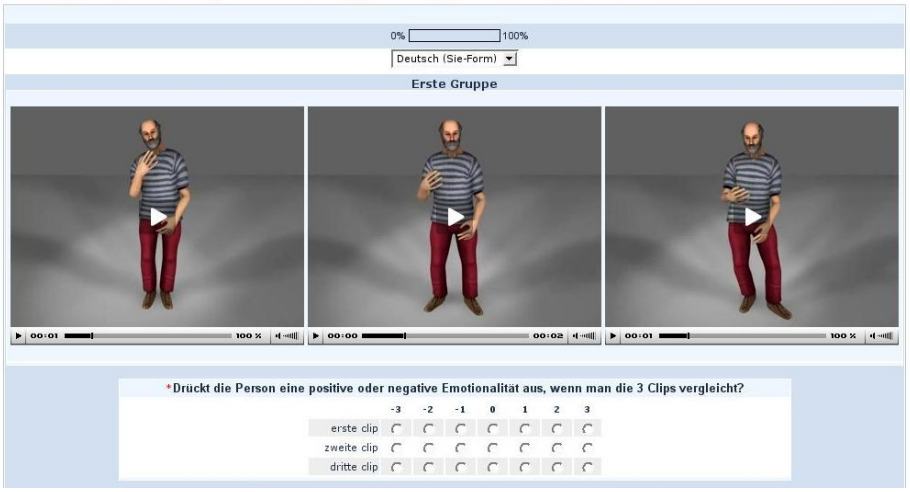

Fig. 2. Screenshot of a form that had to be filled out by a subject

Before starting the experiment, the subject was told to read a short instruction sheet that briefly described the set-up of the experiment and was allowed to ask any questions about the experiment during the test. All the questions relative to the experiment were displayed sequentially. It was not possible for a subject to modify an already answered question, or to browse the questionnaire backwards. The subject was allowed to watch the video as much as he wanted and no time constraint was imposed to finish the questionnaire. The average answering time was 15 minutes.

4.2 Hypothesis

We attempted to verify the following two hypotheses:

[1] Blender is an open source 3D editing and rendering software http://www.blender.org

– (H1): the subjects can discriminate neutral, angry and weary gesture clips along the three quality dimensions: valence, arousal and dominance,
– (H2): the style transformation imitates the angry and weary gesture qualities well.

In order to test these hypotheses, subjects were asked on each question page to rate their perceived amount of valence, arousal or dominance for each clip. Answers were filled with numerical values (integers between -3 and +3). Twice (one time for the *M-W-** material and one time for the *M-T-** material) in the whole questionnaire, the subject was asked to enter an adjective qualifying each video clip in the form of free text input. A questionnaire input form is depicted in figure 4.1. No hints in or around the video clips except gesture quality could help the user determining the actual style (neutral, angry, weary), nor the origin (unfiltered or filtered) of a displayed clip. Placement of the clips (left, center, right) and the total order of clips were randomized.

4.3 Results

The two materials "weather" (*M-W-**) and "train" (*M-T-**) are first compared. For both materials there are three style variants: neutral (*-n*), angry (*-a*) and weary (*-w*). The first question was whether the neutral variants of the two materials are rated similarly on the VAD dimensions (see Figure 3 (a)). We computed an ANOVA with factors material (*M-W-**, *M-W-**) and question (V, A, D). Neither factor material ($F(1,17)$, $p=.21$) nor material-question interaction ($F(2,34)$, $p=.57$) became significant. Therefore, variant n is rated in equally for *M-W-n* and *M-T-n*.

Using the same method we wanted to prove that our transformed motions successfully imitate the angry and weary styles (see Figure 3 (a), middle and right). For *-a*, the ANOVA showed no significance for material ($F(1,17)$, $p=.09$) or for material-question interaction ($F(2,34)$, $p=.13$). Therefore, style *-a* is perceived in the same way for material *M-W* and *M-T*. For style *-w*, a significant effect was found for material ($F(1,17)$, $p<.001$) and material-question interaction ($F(2,34)$, $p<.01$) which means that *M-W* and *M-T* differ.

Since we failed to prove a successful transformation that imitates the weary style *-w*, we compared the generated weary train sequence with the original neutral train sequence. to check whether the transformation did anything significant to the original.Thus, ANOVA was computed with factors style (w, n) and question (V, A, D). The result is inconclusive as the style-question interaction is tendential ($F(2,34)$, $p=.06$). However, *-w* and *-n* do not seem to be too far apart in terms of VAD perception.

Finally we analyzed whether the three style variants were really perceived as being different looking at the VAD dimensions (see Figure 3 (b)). A seperate analysis was made for each material. For each dimension (V, A, D) we computed an ANOVA which proved that the three styles differed. To highlight these differences, a post-hoc Scheffe-test was computer over every dimension pair. For M-W, on the valence (*-V*) dimension we found a significant main effect ($F(2,34)=6.30$,

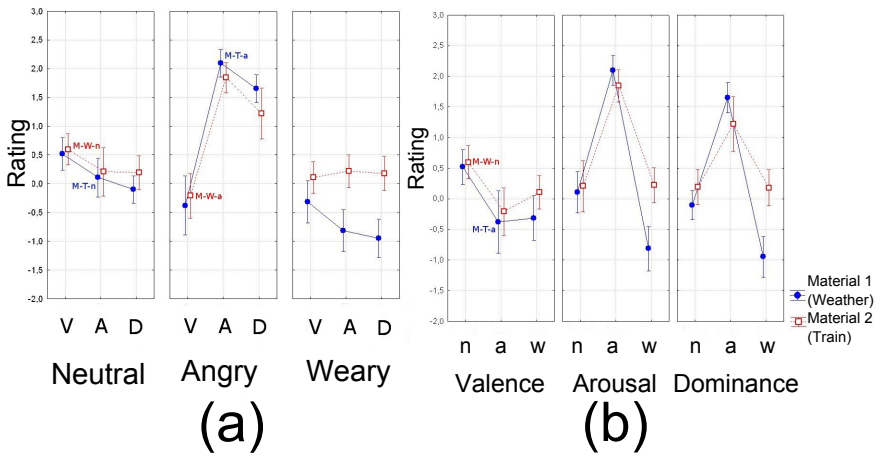

Fig. 3. (a): h1, the subjects can discriminate neutral, angry and weary gesture clips along the dimensions: valence, arousal and dominance.
(b): h2, the style transformation imitates well the angry and weary gesture qualities.

p<.01) which is due to a difference between *-n* and *-a* (p<.05) and between *-n* and *-w* (p<.05) but not due to the difference between *-a* and *-w*. For arousal (*A*) and dominance (*D*) we also found significant main effects (A: F(2,34)=133.74, p<.01; D: F(2,34)=137.27; p<.001) which were due to all three combinations (all p<.001). For material *M-T-**, we found significant main effects for all three questions. On the valence (*V*) dimension (F(2,34)=9.30, p<.001) these were due to differences between *-n* and *-a* (p<.001) and between *-n* and *-w* (p<.05). For both arousal and dominance (A: F(2,34)=49.97, p<.001; D: F(2,34)=18.82, p<.001) they were due to the difference between *-n* and *-a* and between *-a* and *-w* (all p<.001) but not due to the difference between *-n* and *-w*.

5 Discussion

The first hypothesis gives insights on the subjects' ability to discriminate neutral, angry and weary gesture clips along the following three quality dimensions: valence, arousal and dominance.

The results presented show that subjects' ability to separate styled gesture varies regarding the considered dimension. Thus, as stated in the result section and illustrated in the first chart of Figure 3 (b), subjects have difficulties to discriminate the style of motion along the valence dimension. This confirms previous work stating that body motion somehow fails to convey hints about valence [13]. On the contrary, subjects prove to discriminate motion clips much better along the arousal and dominance dimensions (Figure 3 (b), middle chart,

plain curve) and dominance dimension (Figure 3 (b), right chart, plain curve) of the VAD model. This explains why users can better recognise the angry style (high amount of recognizable arousal and dominance) than the weary style (low amount of recognizable arousal and dominance).

The second hypothesis gives us insights on how well style transformation imitates the angry and weary gesture qualities. The results presented in Figure 3 (a), middle chart show that users' perception of original *M-W-a* and transformed angry motion *M-T-a* is very similar. This leads us to believe that our notion transformation method imitates the angry gesture quality well. However, users' perception of original *M-W-w* and transformed *M-T-w* weary motion (Figure 3 (a), right chart) is not perceived in the same way by users. Indeed, results tend to show that weary style transformation does not change much the perceived style of the user compared to the neutral source material (slight tendency for *M-T-n* and *M-T-w* to be perceived the same way, corroborated by the adjective proposed by subjects). We could formulate the following assumptions for explaining the non-recognition of relevant gesture transformation by users. First the temporal retiming was perhaps too slight to be well perceived, thus exagerating this retiming would maybe increase recognition rate. Second, the style transformation we applied does not take into account the stuctural modification that arise between style (hand drop for instance) nor the modification of gesture unit length, (by varying the frequency and length of retractions phases).

6 Conclusion and perspectives

To sum up, this paper presented a method called style transformation. The method consists of transforming an existing motion (either motion captured, manually key-framed or procedurally generated) according to a style that is extracted from an analysis of motion captured files. The resulting motion, although degraded, conveys the subtle characteristics of the reference motion by learning dependencies in data. Besides transfering a specific style, the method can also be used to enhance procedurally synthesized motion that is to date perceived as stiff and unnatural.

An empirical evaluation involving 19 students has then been presented. This evaluation has been designed to verify two issues. First, how well subjects are able to discriminate styled motions along 3 dimension axes of emotion (valence, arousal and dominance). Second, to validate that the tainted motion produced by our style transformation algorithm is perceived as displaying the targeted style.

It has been statistically proved that subjects are able to separate captured body gesture sequences conveying different styles along the VAD model. Also, the gesture sequences produced by our style transformation algorithm have been rated statistically equal as captured styled gesture sequences conveying an angry style. Although the results are not significant when considering weary style, we believe that taking into account gesture unit length and structural modification of gestures would improve the style transformation.

7 Acknowlegdements

We would like to thank Kerstin H. Kipp for her essential contribution in the statistical study. This research has partially been carried out within the framework of the Excellence Cluster Multimodal Computing and Interaction (MMCI), sponsored by the German Research Foundation (DFG). Motion capture Data has been acquired in 2005 in Rennes, France within the framework of the RobEA national program. Sign language gestures have been performed by Alain Cahut.

References

1. Laban, R.: The Mastery of Movement. Northcote House (1988)
2. Wallbot, H.: Bodily expression of emotion. European journal of social psychology **28** (1998) 879–796
3. Chi, D.M., Costa, M., Zhao, L., Badler, N.I.: Emote. In Akeley, K., ed.: Siggraph 2000, Computer Graphics Proceedings, ACM Press (2000) 173–182
4. Hartmann, B., Mancini, M., Pelachaud, C.: Implementing expressive gesture synthesis for embodied conversational agents. In: Gesture in Human-Computer Interaction and Simulation. (2006)
5. Kipp, M., Neff, M., Kipp, K., Albrecht, I.: Toward natural gesture synthesis: Evaluating gesture units in a data-driven approach. In: Proc. of International Conference on Intelligent Virtual Agents. (2006)
6. Neff, M., Kipp, M., Albrecht, I., Seidel, H.P.: Gesture modeling and animation based on a probabilistic recreation of speaker style. ACM Trans. on Graphics (2008, to appear)
7. Schröder, M.: Dimensional emotion representation as a basis for speech synthesis with non-extreme emotions. In: Workshop on Affective Dialogue Systems. (2004)
8. Buisine, S., Hartmann, B., Mancini, M., Pelachaud, C.: Conception et evaluation d'un modèle d'expressivité pour les gestes des agents conversationnels. Revue en Intelligence Artificielle RIA, Special Edition "Interaction Emotionnelle" **20** (2006)
9. Heloir, A., Gibet, S., Multon, F., Courty, N.: Captured motion data processing for real time synthesis of sign language. In: Gesture in Human-Computer Interaction and Simulation. Volume 3881 of LNAI., Berder Island, France, Springer (2005) 168–171
10. Egges, A., Magnenat-Thalmann, N.: Emotional communicative body animation for multiple characters. In: V-Crowds'05. (2005) 31–40
11. Heloir, A., Courty, N., Gibet, S., Multon, F.: Temporal alignment of communicative gesture sequences. Computer Animation and Virtual Worlds **17** (July 2006) 347–357
12. Heloir, A., Gibet, S.: A qualitative and quantitative characterization of style in sign language gestures. In: Gesture in Human-Computer Interaction and Simulation. (2007)
13. Ekman, P., Friesen, W.V.: Hand and Body Cues in the Judgment of Emotion. Perceptual and Motor Skills **24** (1967) 711–724